House Prices and Common Venues in London

1 INTRODUCTION AND DISCUSSION OF BACKGROUND

London is one of the largest cities in Europe with a population of 8,908,081 and population density of 5,666/km2. The city is divided into thirty-two boroughs with the City of London making up a thirty-third borough. Each London borough has a unique demographic and an individual cultural identity. The average residential property values range from £1,294,907 in the borough of Kensington and Chelsea to £303,631 in the borough of Barking and Dagenham. This report will compare boroughs based on average residential property price and common local venues.

The report aims to answer the question: "Can a potential London resident find new boroughs that share characteristics with boroughs that they already enjoy?" Consider a London resident who is looking to move to a new house. This individual may already have an idea of some potential boroughs in which they would like to live. Given the individual's budget and venue preferences, it should be possible to identify other boroughs that share the same characteristics; the new boroughs should be sufficiently similar in residential property value and local venue types. By mapping clusters of boroughs to a chloropleth map of prices, it should be easy to visualise the similarities between boroughs. This will help potential homeowners to identify suitable boroughs in which to search for residential property.

2 DESCRIPTION OF THE DATA

This report will use data from the following sources which are linked in the references section.:

- Government house price data [1]
  This dataset contains the average house price by borough for the year to December 2019. This dataset uses the UK House Price Index (UK HPI) which is calculated by the Office for National Statistics and Land & Property Services Northern Ireland. Data for the UK HPI is provided by HM Land Registry. The dataset is provided in csv format and will be loaded using Pandas.

- Foursquare venue information [2]
  Foursquare users provide information on venues throughout London. This data is hosted by Foursquare and is retrievable through use of the Foursquare API. The required data for this project is the number of venues of a given type in a defined area within each borough.

- Wikipedia co-ordinate data [3]
  This Wikipedia page contains a list of co-ordinates for each London borough. The website will be scraped using the Beautifulsoup package and the latitude and longitude values will be loaded into a Pandas dataframe.

- Geojson file specifying London borough borders [4]
  In order to create a chloropleth map, a geojson file with boundary data of each London borough will be needed. A suitable geojson file is available at the referenced Github page.

3 DATA USE

The co-ordinate data from Wikipedia will be fed into the Foursquare API. This will return a list of the one hundred most common venues within a 1,500 metre radius of the co-ordinate of each London Borough. The k-means clustering algorithm will be used to identify clusters of boroughs that share common venues. The most common venues in each cluster will then be compared to examine the characteristics of the clusters. A short description will be assigned to each cluster.

The house prices will also be visualised and categorised into five bins which will be used to provide to label containing the house price level: low, lower medium, upper medium, high and very high.

The geojson file will provide the borough boundaries for the chloropleth map. The house price data will serve as the variable to create a chloropleth map; the darker shaded regions will represent a higher average house price. The k-means cluster datapoints will be added to the chloropleth map and each datapoint will be labelled. Each label will include: the borough name, a short description of the cluster characteristics, the three most common venues in the borough, the average house price category.

The resulting chloropleth map will serve as a tool for potential homeowner to compare how boroughs with similar common venues compare in terms of average house price. The visualisation may also provide some insight as to the relationship between each cluster and average house prices.

4 METHODOLOGY

I scraped the Wikipedia page for the latitude and longitude values for each London borough and cleaned the data. I used the co-ordinate data to create a folium map to visualise the location of each borough.

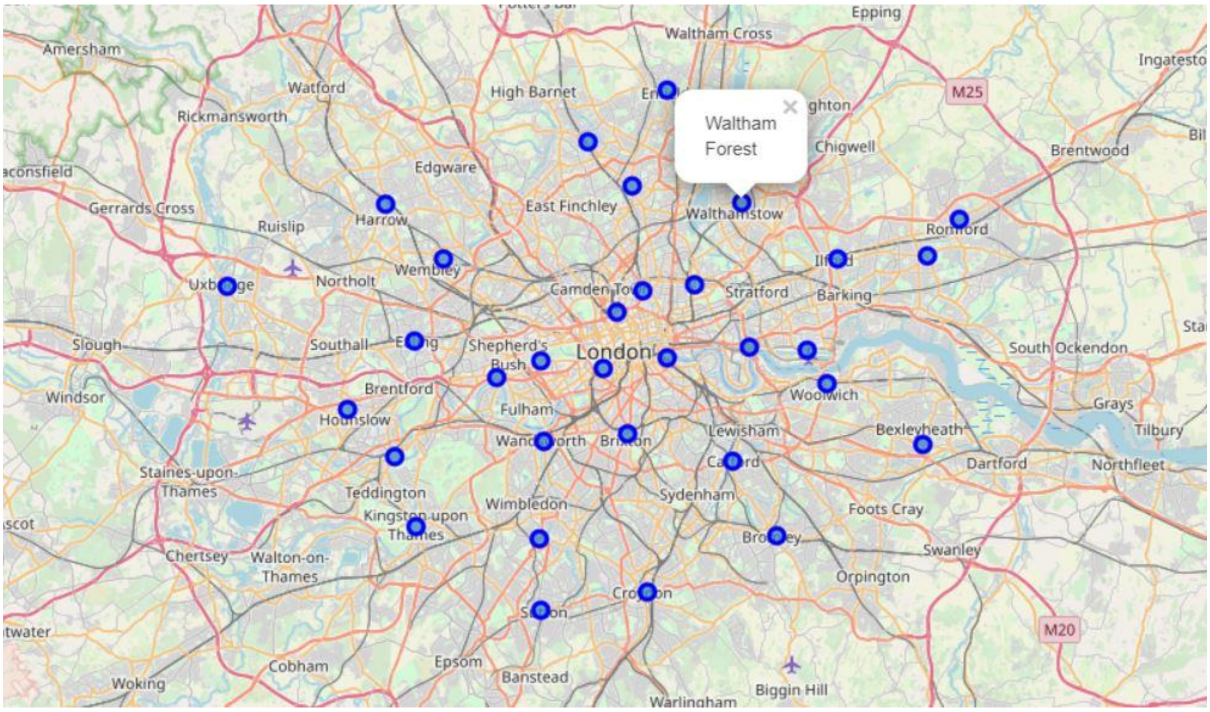


Figure 1: London borough co-ordinates

For this analysis, I decided not to include the City of London for two reasons. The City of London had a significantly lower house sales volume: four sales were recorded in the year to October 2019, whereas each other borough had over one hundred recorded sales. Secondly, the City of London is significantly smaller than the other boroughs at 2.90 km2, therefore the search area for venues may overlap into other boroughs.

I loaded the average house price data from the csv file into a pandas dataframe.

I used the Foursquare API to fetch the top 100 venues within a defined radius from each borough centre. The smallest borough in the dataset is the Borough of Kensington and Chelsea with an area of 12.13 km2. If the borough were a perfect circle, it's radius would be sqrt(12.13/π) = 1960m. As the borough is not a perfect circle, I searched for venues within a 1500m limit. This returned 261 unique categories of venue.

An examination of the number of venues returned shows that some the Foursquare API did not return enough venues for some boroughs. It can be seen from the below table that Barking and Dagenham, Barnet and Merton do not have enough venue references to form a meaningful analysis, so these boroughs were not included.

	Borough
Bexley	60
Newham	60
Hillingdon	53
Harrow	51
Sutton	51
Redbridge	50
Lewisham	47
Merton	37
Barnet	21
Barking and Dagenham	19

Figure 2: Boroughs with lowest venues returned

I then compared the top ten venues in each borough.

	Borough	**1st Most Common Venue**	**2nd Most Common Venue**	**3rd Most Common Venue**	**4th Most Common Venue**	**5th Most Common Venue**	**6th Most Common Venue**	**7th Most Common Venue**	**8th Most Common Venue**	**9th Most Common Venue**	**10th Most Common Venue**
0	Bexley	Pub	Fast Food Restaurant	Coffee Shop	Clothing Store	Supermarket	Italian Restaurant	Grocery Store	Hotel	American Restaurant	Pharmacy
1	Brent	Coffee Shop	Hotel	Clothing Store	Bar	Indian Restaurant	Sandwich Place	Grocery Store	Italian Restaurant	Supermarket	Pizza Place
2	Bromley	Pub	Coffee Shop	Clothing Store	Supermarket	Pizza Place	Gym / Fitness Center	Indian Restaurant	Park	Electronics Store	Bar
3	Camden	Coffee Shop	Café	Pizza Place	Hotel	Breakfast Spot	History Museum	Exhibit	Burger Joint	Bookstore	Beer Bar
4	Croydon	Pub	Coffee Shop	Clothing Store	Mediterranean Restaurant	Park	Indian Restaurant	Bookstore	Italian Restaurant	Portuguese Restaurant	Sushi Restaurant

Figure 3: Top ten venues in each borough (first four entries)

4.1 MODELLING

I used the k-means clustering algorithm to cluster the boroughs into four clusters. The optimal value for k was determined using the 'elbow method'. As can be seen from the figure 4 graph, the optimal value for k is three: the 'elbow' point on the graph.

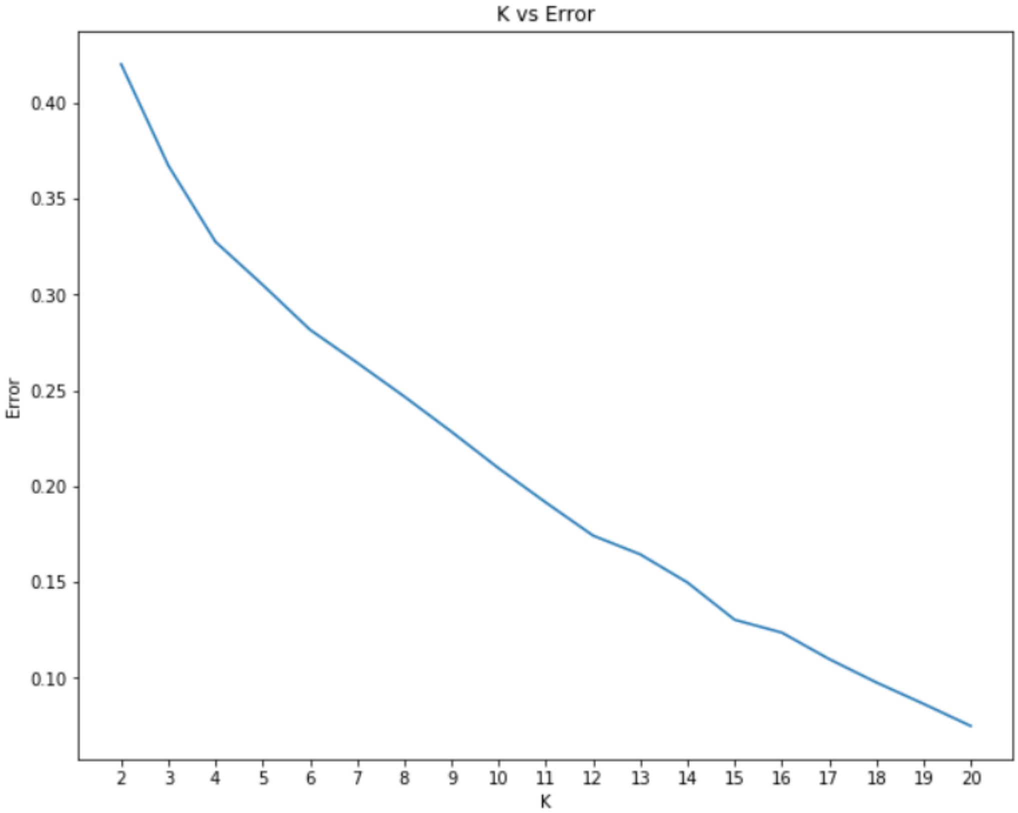


Figure 4: Model error for different values of k

In order to give some insight into the characteristics of the clusters, I estimated the most common within each cluster. The two pivot tables below show the first and second most common venues respectively in each cluster.

Cluster Label	Café	Clothing Store	Coffee Shop	Grocery Store	Hotel	Indian Restaurant	Pub
0	1	0	1	0	0	0	8
1	0	0	2	0	0	1	0
2	0	1	2	1	0	0	7
3	0	0	3	0	2	0	0

Figure 5: Pivot table of the first most common venues by cluster

Cluster Label	Burger Joint	Café	Clothing Store	Coffee Shop	Fast Food Restaurant	Grocery Store	Hotel	Indian Restaurant	Pizza Place	Pub	Supermarket
0	0	3	0	4	0	0	0	0	1	2	0
1	0	0	0	0	0	0	2	1	0	0	0
2	0	0	1	5	1	2	0	0	0	1	1
3	1	1	0	2	0	0	1	0	0	0	0

Figure 6: Pivot table of the second most common venues by cluster

The first and second most common venues in the above pivot tables can be more easily visualised in bar graphs.

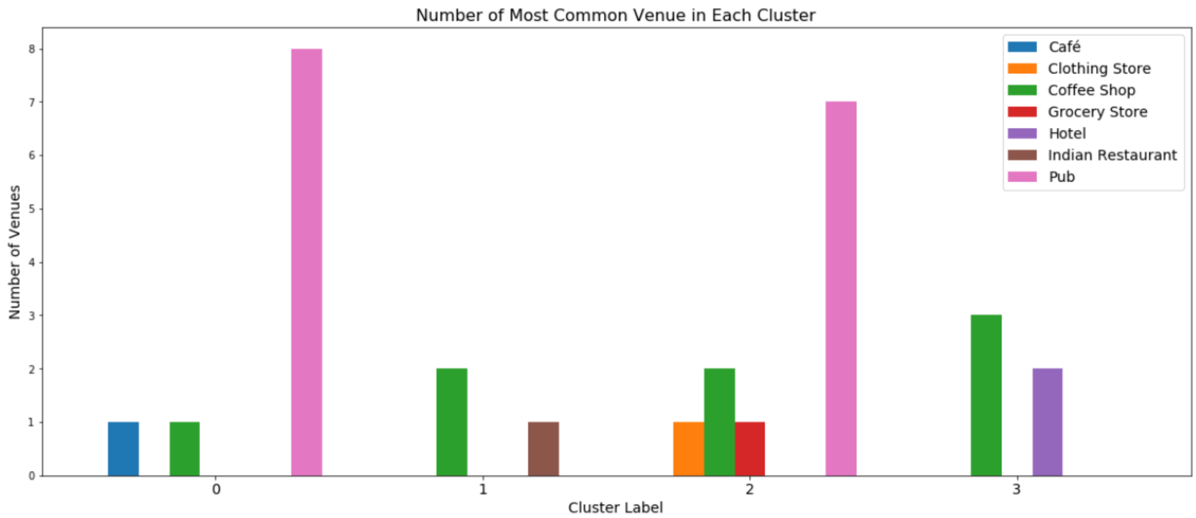


Figure 7: Bar graph of the first most common venues by cluster

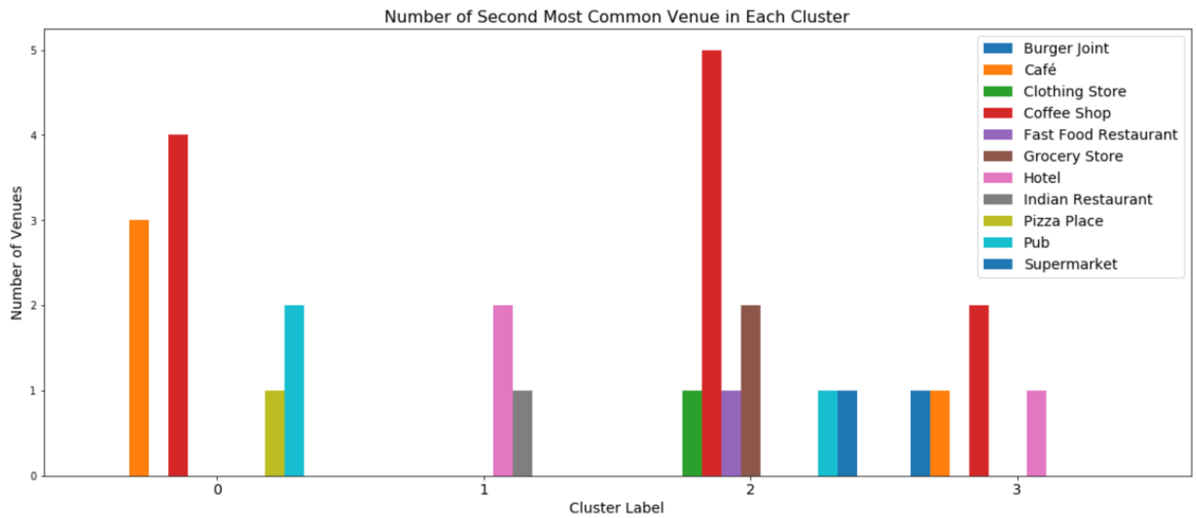


Figure 8: Bar graph of the second most common venues by cluster

I assigned a descriptive label to each cluster based on the most common venues in each:

- Cluster 0: Pubs and cafes.
- Cluster 1: Indian restaurants and cafes/hotels.
- Cluster 2: Pubs, cafes, food and clothes.
- Cluster 3: Coffee and hotels.

It should be noted that 'cafés' and 'coffee shops' are returned as separate venue types. As these venues are very similar, clusters with high numbers of coffee shops have been treated as if they have a high number of cafés and vice-versa.

The clusters can visualised by adding them to the folium map of London.

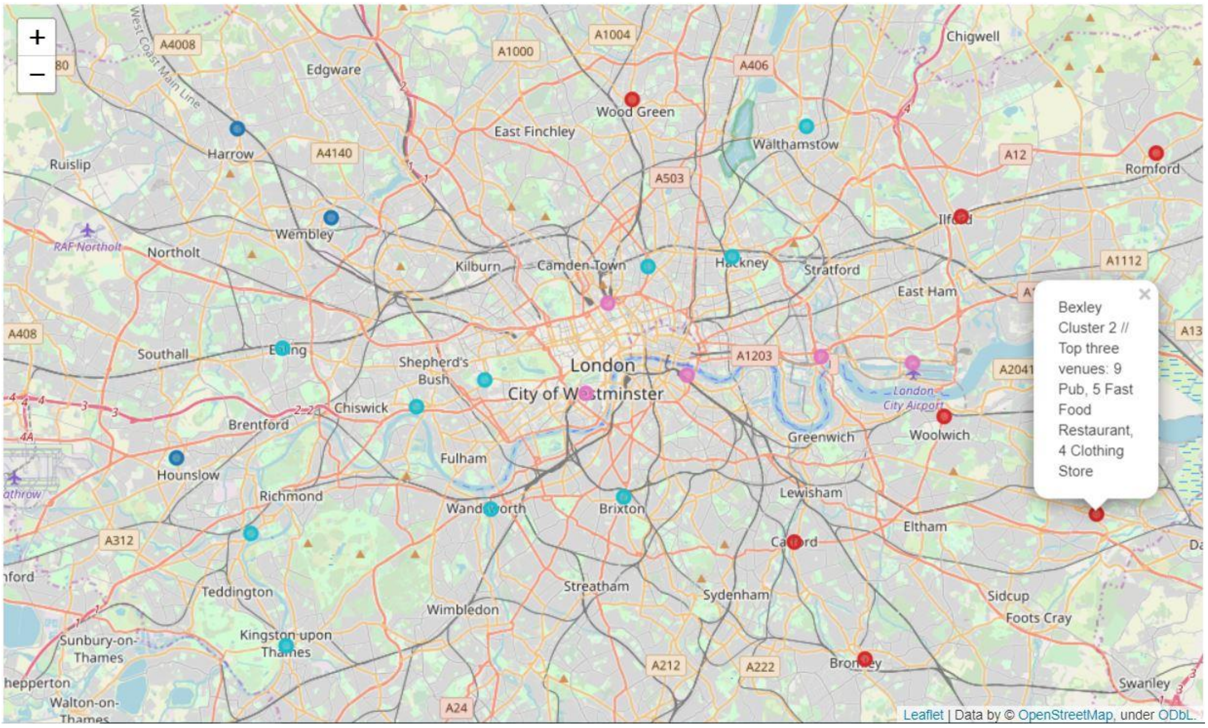


Figure 9: Map of clusters

I created a dataframe that contained the number and type of the three most common venues for each borough. This information will be shown on the interactive map.

	Borough	Top 3 Venues
0	Bexley	9 Pub, 5 Fast Food Restaurant, 4 Clothing Store
1	Brent	8 Coffee Shop, 7 Clothing Store, 7 Hotel
2	Bromley	6 Pub, 5 Coffee Shop, 4 Clothing Store
3	Camden	10 Coffee Shop, 6 Café, 5 Hotel
4	Croydon	10 Pub, 9 Coffee Shop, 6 Clothing Store

Figure 10: Top three venues for each borough (first four entries)

I created a histogram with five bins; we can see that most of boroughs share a lower average house price.

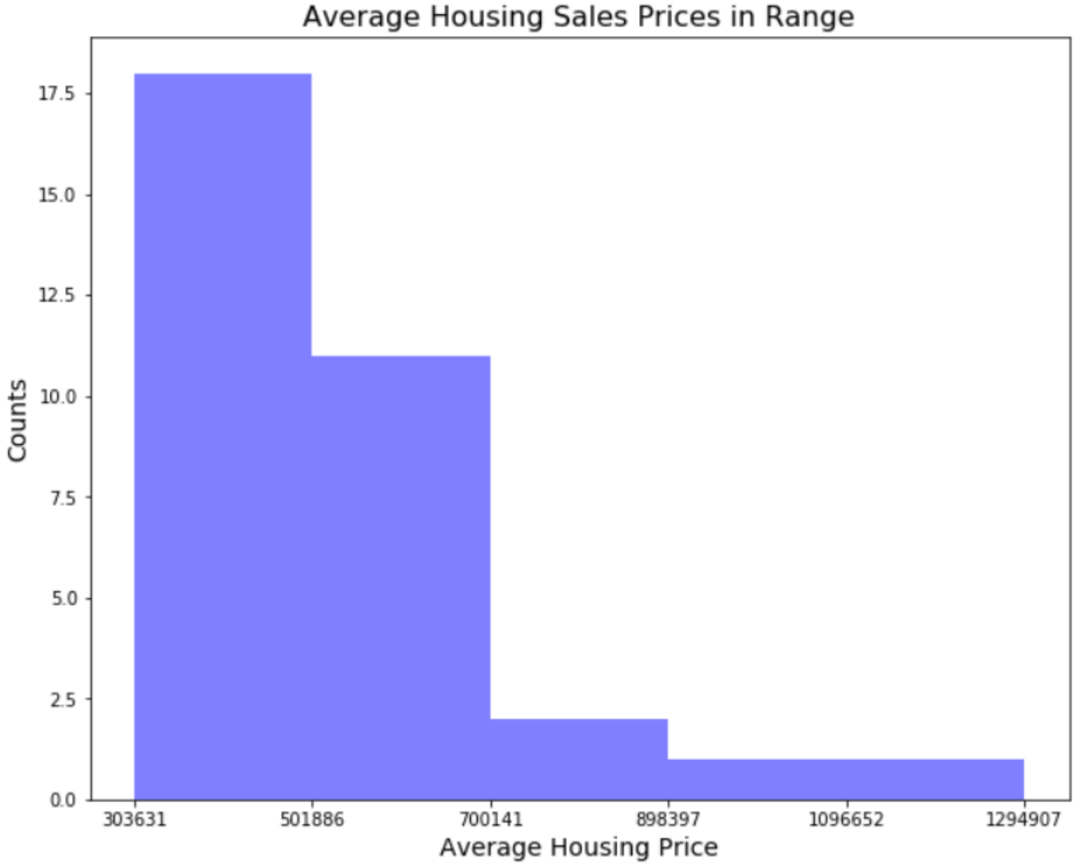


Figure 11: Histogram of borough average house prices

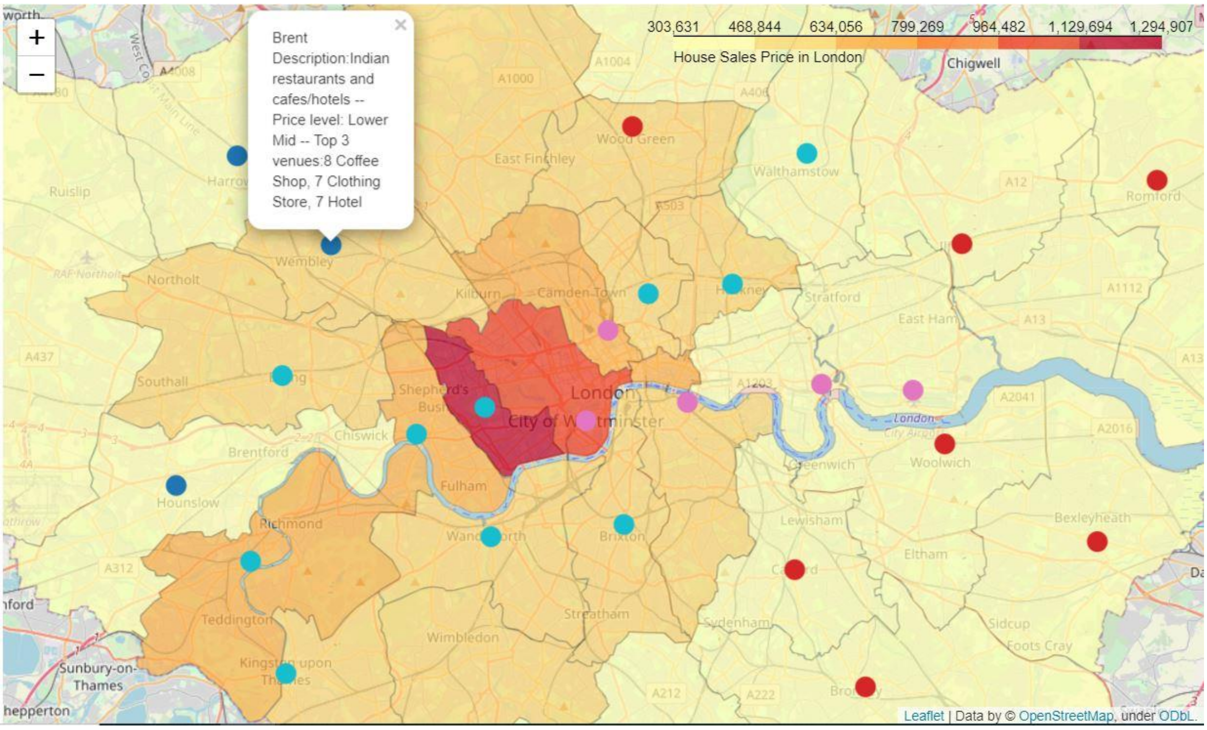
I used the histogram bins to create five categories of house price: low, lower mid, upper mid, high and very high. For example, boroughs with an average price between £303,631 and £501,886 will be categorised as having a 'low' price level. I now have all the labels that will be used for the folium map popups:

	Borough	Latitude	Longitude	Top 3 Venues	Price Level	Cluster Labels	Description
0	Bexley	51.4549	0.1505	9 Pub, 5 Fast Food Restaurant, 4 Clothing Store	Low	2	Pubs, cafes, food and clothes
1	Brent	51.5588	-0.2810	8 Coffee Shop, 7 Clothing Store, 7 Hotel	Lower Mid	1	Indian restaurants and cafes/hotels
2	Bromley	51.4039	0.0198	6 Pub, 5 Coffee Shop, 4 Clothing Store	Low	2	Pubs, cafes, food and clothes
3	Camden	51.5290	-0.1250	10 Coffee Shop, 6 Café, 5 Hotel	Upper Mid	3	Cafes and hotels
4	Croydon	51.3714	-0.0970	10 Pub, 9 Coffee Shop, 6 Clothing Store	Low	2	Pubs, cafes, food and clothes

Figure 12: table with map labels (first four entries)

5 RESULTS

To analyse the findings, the borough clusters have been superimposed onto an interactive chloropleth of average house prices. As can be seen in the legend, the boroughs shaded darker red have a higher average house price. Each borough has a colour label corresponding to its respective cluster. By clicking on a cluster, a popup appears with the borough name, a description of the characteristics of its cluster, the house price category, and the top three most common venues in the borough.



From the above map, we can gain some insights about the nature of the clusters regarding where they are located geographically and whether individual boroughs within the clusters share similar average house prices.

6 Discussion

Cluster 0: 'pubs and cafes' is represented by the light blue markers. We can see that this cluster tends to contain boroughs with a higher average price. This cluster also has a moderate number of restaurants. The boroughs within this cluster tend to be located in the south-west and north-east of London.

Cluster 1: 'Indian restaurants and cafes/hotels' is represented by the dark blue markers. This cluster contains the lowest number of boroughs which makes it difficult to make meaningful inferences. The majority of boroughs in this cluster tend to have a lower average house price and are located in the west of the city.

Cluster 2: 'Pubs, cafes, food and clothes' is represented by the red markers. This cluster contains a mix of different types of venues: some include food/drink and various shopping establishments. The boroughs within this cluster are characterised as having a low average house price and are located in the west of the city.

Cluster 3: 'Coffee and hotels' is represented by the pink markers. Interestingly, this cluster contains many boroughs that are located on the river. This may be partially due to the popularity of opening hotels on the riverfront. This cluster appears to have the highest variation in average house price, ranging from low to high. The boroughs within this cluster tend to be located to the east of the city.

7 Conclusion

In this study, I grouped the boroughs of London into clusters with similar common venues and compared the characteristics of the resulting clusters. I mapped the clusters on a chloropleth map of average house prices to determine whether boroughs within each cluster shared similar average house prices. I found that boroughs within the same cluster often shared geographical characteristics and had a relatively similar average house price.

It is interesting that some boroughs within some clusters shared price and location characteristics: this information was not used in the k-means clustering algorithm which suggests that the type of common venues in a borough may be used to estimate the average house price. An interesting area for further study would be to determine the exact relationship between common venues and average house prices. This information could be used to predict changes in house prices, given a shift in the popularity of new types of business opening.

8 References

- [1] London house price data
- [2] Foursquare API
- [3] Wikipedia co-ordinate data
- [4] Geojson file specifying London borough borders